Valverde-Albacete, F. J. & Peláez-Moreno, C. (2017). The evaluation of data sources using multivariate entropy tools. *Expert Systems with Applications*, vol. 78, pp. 145–157.

DOI: 10.1016/j.eswa.2017.02.010

The Evaluation of Data Sources using Multivariate Entropy Tools

Francisco J. Valverde-Albacete[a], Carmen Peláez-Moreno[a,*]

[a]*Departamento de Teoría de la Señal y Comunicaciones, Universidad Carlos III de Madrid, 28911 Leganés, Spain*

Abstract

We introduce from first principles an analysis of the information content of multivariate distributions as information sources. Specifically, we generalize a balance equation and a visualization device, the Entropy Triangle, for multivariate distributions and find notable differences with similar analyses done on joint distributions as models of information channels.

As an application, we extend a framework for the analysis of classifiers to also encompass the analysis of data sets. With such tools we analyze a handful of UCI machine learning task to start addressing the question of how well do datasets convey the information they are supposed to capture about the phenomena they stand for.

Keywords: Machine Learning Evaluation, Dataset Entropy, Multivariate Entropy, Entropic Measures, Exploratory Analysis, Entropy Ternary Diagram, Entropy Balance Equation.

1. Introduction and Motivation

In this paper we introduce an information-theoretic perspective into the problem of characterizing the datasets in machine learning tasks, and obtain several tools, both theoretical and practical, to explore such problem.

*Corresponding author

Email addresses: fva@tsc.uc3m.es (Francisco J. Valverde-Albacete), carmen@tsc.uc3m.es (Carmen Peláez-Moreno)

Information-theory was founded by Shannon in his two-part seminal paper (Shannon, 1948a,b) to provide a mathematical background to the transmission of information in the presence of noise. The last 60 years of engineering practice have revealed that this setting is far broader than initially envisaged, and many problems, both theoretical and applied, can be characterized as "relating to the transmission of information", that is, in information-theoretical terms (see, e.g. MacKay, 2003; Brillouin, 1962).

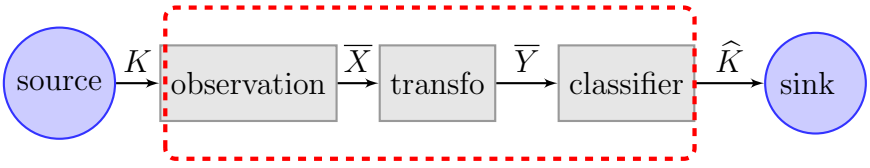
In particular, a strong current to use information-theoretic principles and heuristics in machine learning (Principe, 2010) and statistical inference (Jaynes, 1996, Chap. 11), and several methods for evaluation and analysis based on entropic measures with diverse applications have been recently published (Valverde-Albacete & Peláez-Moreno, 2010; Zhou et al., 2013; Rödder et al., 2014; Chen et al., 2014; Valverde-Albacete & Peláez-Moreno, 2014; Hempelmann et al., 2016).

As early as (McGill, 1954), there emerged an interest in better understanding how the transmission of information in the *multivariate setting*—that is, among multiple variables—compares to the *bivariate setting* used by Shannon for variables X and Y. For the purpose at hand consider the scheme of Figure 1.(a) conceptualizing the supervised machine learning task of multi-class classification, cast in an information-theoretic setting. There is a set of m realizations of a random vector $\overline{X}$ of *(observed) variables or features* paired with as many realizations of a *class variable* K. The set of pairs of instances $\{(k^i, \overline{x}^i)\}_{1\leq i\leq m}$ will be called a *dataset*. For unsupervised tasks, we typically ignore or disregard K.

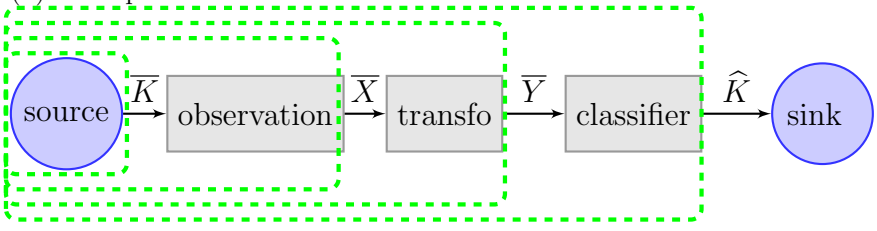
The feature instances $\overline{X} = \overline{x}^i$ may be further transformed to obtain instances of a random vector $\overline{Y}$, through a tranformation function $f : \overline{X} \to \overline{Y}, \overline{x}^i \mapsto \overline{y}^i = f(\overline{x}^i)$ with desired characteristics, e.g. statistical independence among the transformed features. For supervised classification, classifier induction is the subtask of inducing a function $k : \overline{Y} \to K, \overline{y}^i \mapsto \hat{k}^i = k(\overline{y}^i)$ that tries to estimate the original K but can only obtain the estimate $\hat{K}$.

For an end-to-end measure of the effectiveness of this procedure of estimating $\widehat{K}$ from K as per the box in Figure 1.(b), a Shannon-type equation on the entropies around a bivariate joint distribution was introduced in (Valverde-Albacete & Peláez-Moreno, 2010) and later refined in (Valverde-Albacete & Peláez-Moreno, 2014) (see Section 2.1). It was named the *balance equation* and it leads to a new kind of exploratory graph for entropies: a ternary or *de Finetti* diagram of entropies, also called the *entropy triangle (ET)* (see Section 2.2). Both tools have been used to evaluate multiclass clas-

(a) Simplified multiclass classification scheme

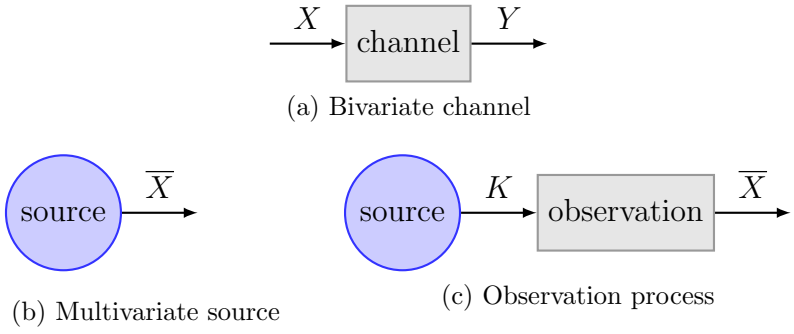


(b) Conceptual measurement scheme for end-to-end evaluation



(c) Conceptual measurement schemes for the information content of sources

Figure 1: **Schematic representation of a multi-class classification task and measurement schemes for information-theoretic quantities.**



(a) Bivariate channel

(b) Multivariate source

(c) Observation process

Figure 2: **Examples of systems susceptible of analysis with the techniques discussed in the paper**. (a) Single-input single output system studied with previous entropy triangles, and (b) opaque multivariate source, (c) multivariate source coming from an observation process, to be studied with the techniques presented in this paper.

sifiers (Valverde-Albacete et al., 2013) using the joint distribution of results implicit in the confusion matrix over the classified instances as evaluated on the train and test data (Theodoridis & Koutroumbas, 2006; Murphy, 2012) (see Section 2.3).

Again, such tools, allow us to analyze single-input single-output processing blocks like that in Figure 2.(a). But in this paper, we would like to investigate whether there are analogous results for multivariate stochastic sources of information whose block diagram fragment in focus is that of Figure 2.(b). For that purpose, let $\overline{X} = \{X_i \mid 1 \leq i \leq n\}$ be a set of discrete random variables with joint multivariate distribution $P_{\overline{X}}(\overline{x}) = P_{X_1 \ldots X_n}(x_1 \ldots x_n)$— where $\overline{x} = x_1 \ldots x_n$ is a tuple of n elements—with marginals $P_{X_i}(x_i) = \sum_{j \neq i} P_{\overline{X}}(\overline{x})$.

But we would also like to study the related procedure of observing a random variable K through an observation process whose result is the random vector $\overline{X}$, as depicted in Figure 2.(c), which is precisely the setting of supervised classification. With this goal in mind, in supervised tasks we may select one of the variables to represent a *class index* K in this (categorical or discrete) setting. When the support of K has more than two values $|\text{supp}(K)| \geq 2$ we call this setting *multiclass classification*; if $|\text{supp}(K)| = 2$, we call it *(binary) classification*. When this is the model of the data (as in Section 4.3) we will suppose that the classification variable K is actually adjoined to variable vector $\overline{X}$ and it is interpreted as the underlying process captured by the observation data.

In the following, we first review in Section 2 the theory and methods behind the balance equation and the entropy triangle, including a discussion

of the issues that need to be addressed for their multivariate generalization, and ending with a set of problems that have to be solved in order to do so. In Section 3 we present our main theoretical contribution, the generalizations of the balance equation and the entropy triangle for multivariate distributions, and in Section 4 we introduce examples of uses for these tools for the exploratory analysis of machine learning tasks, both supervised and unsupervised. We end with a brief discussion of alternate representation mechanisms for entropy balances, the uses of such tools and some conclusions.

2. Methods and Tools

2.1. The joint entropy balance of two variables

The tools we propose are based on an often overlooked decomposition of the joint entropy of two random variables (Valverde-Albacete & Peláez-Moreno, 2010). Figure 3 depicts this decomposition showing the three crucial regions:

- The *divergence with respect to uniformity*, $\Delta H_{P_X \cdot P_Y}$, between the joint distribution where P_X and P_Y are independent and the uniform distributions U_X and U_Y with the same cardinality of events as P_X and P_Y .

$$\Delta H_{P_X \cdot P_Y} = H_{U_X \cdot U_Y} - H_{P_X \cdot P_Y} \quad .$$

- The *mutual information*, $MI_{P_{XY}}$, quantifies the force of the stochastic binding between P_X and P_Y .

$$MI_{P_{XY}} = H_{P_X \cdot P_Y} - H_{P_{XY}}$$

- The *variation of information*, $VI_{P_{XY}}$, embodies the residual entropy, not used in binding the variables.

$$VI_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}}$$

Each of these quantities provide intuitions into the behavior of P_X, P_Y and P_{XY} used to advantage in applications (cfr. Section 2.3), and we would like to reproduce them in a multivariate setting for applications like feature filtering (Brown et al., 2012) or multi-label classification (Gibaja & Ventura, 2015).

Note that all of these quantities are positive. In fact from the previous decomposition the following *balance equation* is evident,

$$H_{U_X \cdot U_Y} = \Delta H_{P_X \cdot P_Y} + 2 * MI_{P_{XY}} + VI_{P_{XY}} \tag{1}$$
$$0 \leq \Delta H_{P_X \cdot P_Y}, MI_{P_{XY}}, VI_{P_{XY}} \leq H_{U_X \cdot U_Y}$$

where the bounds are easily obtained from distributional considerations (Valverde-Albacete & Peláez-Moreno, 2010).

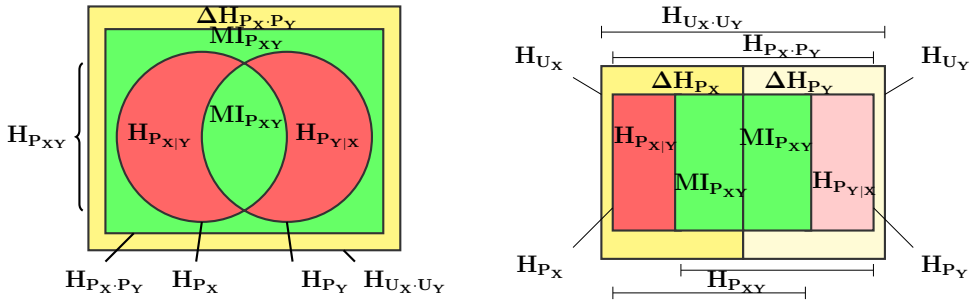


Figure 3: **(Color online) Extended entropy diagrams related to a bivariate distribution, from (Valverde-Albacete & Peláez-Moreno, 2010).** The bounding rectangle is the joint entropy of two uniform (hence independent) distributions U_X and U_Y of the same cardinality as input probability distribution P_X and output P_Y, resp. The expected mutual information $MI_{P_{XY}}$ appears *twice* in (a) and this makes the diagram split for each variable symmetrically in (b).

2.2. From the balance equation to the joint entropy triangle

If we normalize (1) by the overall entropy $H_{U_X \cdot U_Y}$ we obtain

$$1 = \Delta' H_{P_X \cdot P_Y} + 2 * MI'_{P_{XY}} + VI'_{P_{XY}} \tag{2}$$
$$0 \leq \Delta' H_{P_X \cdot P_Y}, MI'_{P_{XY}}, VI'_{P_{XY}} \leq 1$$

Equation (2) is the 2-simplex in normalized $\Delta H'_{P_X \cdot P_Y} \times 2MI'_{P_{XY}} \times VI'_{P_{XY}}$ space. Each joint distribution P_{XY} can be characterized by its joint entropy proportions, or entropic *composition* (Aitchison, 1982; Pawlowsky-Glahn et al., 2015) $F(P_{XY}) = [\Delta H'_{P_{XY}}, 2 \times MI'_{P_{XY}}, VI'_{P_{XY}}]$. Its projection onto the plane with director vector $(1, 1, 1)$ is its *de Finetti (entropy) diagram*, represented in Fig. 4 which shows as an equilateral triangle, hence the alternative name of *entropy triangle*.

Therefore, *every binary distribution shows as a point in the triangle and the position in the triangle entails qualities of the distribution:*

- The lower side of the triangle is the geometric locus of distributions with independent marginals: if $P_{XY} = P_X \cdot P_Y$ then $F(P_{XY}) = [\cdot, 0, \cdot]$.
- The left side is the geometric locus of distributions with uniform marginals. If $P_X = U_X$ and $P_Y = U_Y$ then $F(P_{XY}) = [0, \cdot, \cdot]$.
- Finally, the right-hand side is the locus of distributions with identical marginals: if $P_X = P_Y$—that is, $H_{P_X} = H_{P_Y} = MI_{P_{XY}}$—then $F(P_{XY}) = [\cdot, \cdot, 0]$.

2.3. Application: evaluating classifiers

The evaluation of classifiers is fairly simple using their confusion matrices and the schematic in Fig. 4.

1. Classifiers on the bottom side of the triangle *transmit no mutual information* from input to output: they have not profited from being exposed to the data.
2. Classifiers on the right hand side have diagonal confusion matrices, hence *perfect (standard) accuracy*.
3. Classifiers on the left hand side operate on perfectly balanced data distributions, hence they are *solving the most difficult multiclass problem* (from the point of view of an uninformed decision).

Of course, combinations of these conditions provide specific kinds of classifiers. Those at the apex or close to it are obtaining the highest accuracy possible on very balanced datasets and transmitting a lot of mutual information hence they are the *best classifiers* possible. Those at or close to the left vertex are essentially not doing any job on very difficult data: they are *the worst classifiers*. Those at or close to the right vertex are not doing any job on very easy data for which they claim to have very high accuracy: they are *specialized (majority) classifier*s and our intuition is that they are the kind of classifiers that generate the accuracy paradox, whereby classifiers with higher test set accuracy provide lower deployment accuracy, since the data in the deployment scenario might not be as imbalanced as in lab conditions (Valverde Albacete & Peláez-Moreno, 2016).

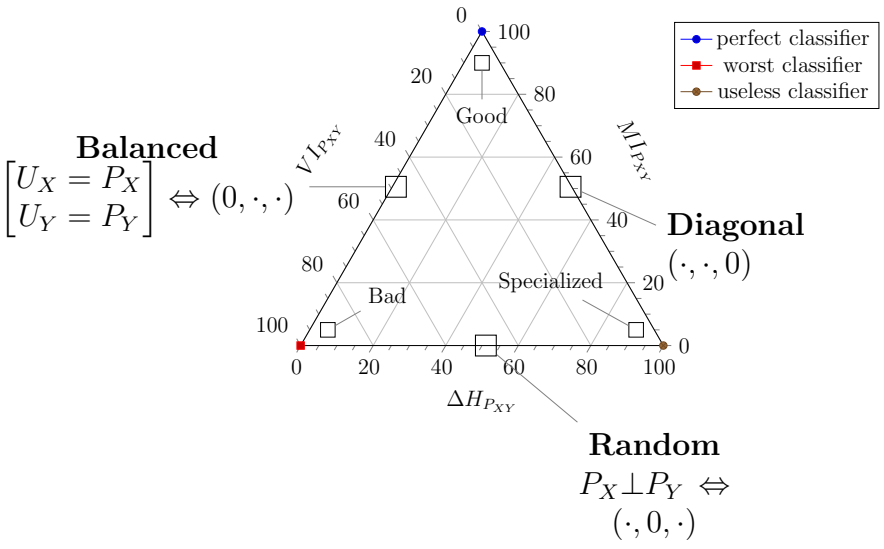


Figure 4: **(color online) Schematic Entropy Triangle showing interpretable zones and extreme cases of classifiers, from (Valverde-Albacete & Peláez-Moreno, 2010).** The annotations on the center of each side are meant to hold for that whole side.

In just this guise, the ET has already been successfully used in the evaluation of Speech Recognition systems (Valverde-Albacete & Peláez-Moreno, 2010; Mejía-Navarrete et al., 2011), sentiment analysis (Valverde-Albacete et al., 2013), and other classification tasks, using one of several implementations in Matlab (Valverde-Albacete & Peláez-Moreno, 2010), R (Valverde-Albacete, 2016) and as a Weka plugin[1].

2.4. The split balance equation and entropy triangle of two variables

In (Valverde-Albacete & Peláez-Moreno, 2010) it is reasoned how (1) may be split into two equations. Briefly, since both U_X and U_Y on the one hand and P_X and P_Y are independent as marginals of U_{XY} and $P_X P_Y$, respectively, we may write:

$$\Delta H_{P_X P_Y} = (H_{U_X} - H_{P_X}) + (H_{U_Y} - H_{P_Y}) = \Delta H_{P_X} + \Delta H_{P_Y} \tag{3}$$

where

$$\Delta H_{P_X} = H_{U_X} - H_{P_X} \qquad \Delta H_{P_Y} = H_{U_Y} - H_{P_Y} \tag{4}$$

[1] `http://apastor.github.io/entropy-triangle-weka-package`

This and the occurrence of twice the expected mutual information in Eq. (1) suggests a different information diagram, depicted in Fig. 3(b). Both variables X and Y now appear somehow decoupled—in the sense that the areas representing them are disjoint—yet there is a strong coupling in that the expected mutual information appears in both H_{P_X} and H_{P_Y} .

$$H_{U_X} = \Delta H_{P_X} + MI_{P_{XY}} + H_{P_{X|Y}} \qquad H_{U_Y} = \Delta H_{P_Y} + MI_{P_{XY}} + H_{P_{Y|X}} \,. \quad (5)$$

The *split entropy triangle* focuses on these quantities *for each component variable or marginal distribution in the joint distribution.* They describe the *marginal fractions* of entropy when the normalization is done with H_{U_X} and H_{U_Y} respectively

$$\begin{aligned} F_X(P_{XY}) &= [\Delta H'_{P_X}, MI'_{P_{XY}}, VI'_X = H'_{P_{X|Y}}] \\ F_Y(P_{XY}) &= [\Delta H'_{P_Y}, MI'_{P_{XY}}, VI'_Y = H'_{P_{Y|X}}] \end{aligned} \quad (6)$$

hence we may consider the de Finetti marginal entropy diagrams for both F_X and F_Y to visualize the entropy changes from input to output.

2.5. Related work: multivariate generalizations of Mutual Information

In order to pave the way for a discussion of problems in Section 2.6, we next review the different "flavors" of information measures describing sets of more than two variables. First, we would like to note that information diagrams (I-diagrams) (Reza, 1961)—such as those of Figs. 3 and 5—are a powerful tool to visualize the interaction of distributions in the bivariate case, but the following caveats apply:

- Their multivariate generalization is only warranted when signed measures of probability are considered, since it is well-known that some of these "areas" can be *negative*, contrary to geometric intuitions on this respect.

  The differences with the bivariate case concentrate in the green areas of Figure 5. The case illustrated is $n = 3$ since higher orders are difficult to visualize in this guise (see, e.g. (James et al., 2011) for the case where n=4).

- We should not forget about the bounding rectangles that appear when considering the most entropic distributions with similar support to the ones being graphed (Valverde-Albacete & Peláez-Moreno, 2010). This is the sense of the bounding rectangles in Figures 3.(a) and 5.

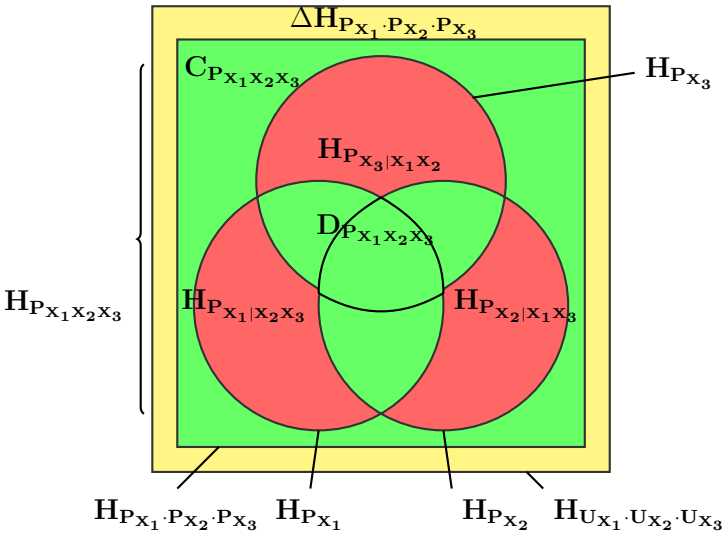


Figure 5: **(Color online) Extended entropy diagram of a trivariate distribution.** The bounding rectangle is the joint entropy of uniform (hence independent) distributions U_{X_i} of the same cardinality as distribution P_{X_i}. The green area is the sum of the multi-information (total correlation) $C_{P_{\overline{X}}}$ and the dual total correlation $D_{P_{\overline{X}}}$.

A recent review on multivariate information measures is (James et al., 2011). An interesting methodological point made there is to call *information* those measures which involve amounts of entropy shared by multiple variables and *entropies* those that do not.[2] For instance, from first principles we must consider the fact that every random variable has a residual entropy which might not be explained away by the information provided by the other variables. $H_{P_{X_i|X_i^c}}$ where $X_i^c = \overline{X} \setminus \{X_i\}$. We call *residual information* (James et al., 2011) or *(multivariate) variation of information* (Meila, 2007; Valverde Albacete & Peláez-Moreno, 2016) a generalization of the same quantity in the bivariate case, the sum of these quantities across the set of random variables:

$$VI_{P_{\overline{X}}} = \sum_{i=1}^{n} H_{P_{X_i|X_i^c}} \,. \tag{7}$$

[2] Although this certainly poses a conundrum for the entropy written as the self information $H_{P_X} = MI_{P_X P_X}$.

James et al. (2011) also point out that some of the information measures stem from focusing in a particular property of the bivariate mutual information and generalize it to the multivariate setting. The properties in question are:

$$MI_{P_{XY}} = H_{P_X} + H_{P_Y} - H_{P_{XY}} \tag{8}$$

$$MI_{P_{XY}} = H_{P_X} - H_{P_{X|Y}} = H_{P_Y} - H_{P_{Y|X}} \tag{9}$$

$$MI_{P_{XY}} = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \tag{10}$$

- the *total correlation* (Watanabe, 1960), *integration* (Tononi et al., 1994) or *multiinformation* (Studený & Vejnarová, 1998) is a generalization of (8), represented by the green area outside $H_{P_{\overline{X}}}$.

$$C_{P_{\overline{X}}} = H_{\Pi_{\overline{X}}} - H_{P_{\overline{X}}} \tag{11}$$

- the *dual total correlation* (Han, 1978; Abdallah & Plumbley, 2012) or *interaction complexity* (Tononi, 1998) is a generalization of (9), represented by the green area inside $H_{P_{\overline{X}}}$

$$D_{P_{\overline{X}}} = H_{P_{\overline{X}}} - VI_{P_{\overline{X}}} \tag{12}$$

- the *interaction information* (McGill, 1954), *multivariate mutual information* (Sun Han, 1980) or *co-information* (Bell, 2003) is the generalization of (10), the total amount of information to which all variables contribute.

$$MI_{P_{\overline{X}}} = \sum P_{\overline{X}}(\overline{x}) \log \frac{P_{\overline{X}}(\overline{x})}{\Pi_{\overline{X}}(\overline{x})} \tag{13}$$

It is represented by the inner convex area (within the dual total correlation), but note that it may in fact be negative for $n > 2$ (Abdallah & Plumbley, 2010).

- the *local exogenous information* (James et al., 2011) or the *bound information* (Valverde Albacete & Peláez-Moreno, 2016) is the addition of the total correlation and the dual total correlation

$$M_{P_{\overline{X}}} = C_{P_{\overline{X}}} + D_{P_{\overline{X}}}\,. \tag{14}$$

2.6. A roadmap for generalization

Given the previous consideration, our next step will be to try and generalize the balance equations and entropy triangles to the case of multiple variables $\overline{X}$ taken as a joint source of information.

In a nutshell, given a random vector $\overline{X}$ distributed after the multivariate distribution $\overline{X} \sim P_{\overline{X}}$, the above sections suggest the following manner of proceeding:

- Find an analogue of the balance equation related to the entropies of $P_{\overline{X}}$.
- Transform the balance equation into a normalized equation for a 2-simplex, and give intuitive meanings to each coordinate.
- Split the balance equation adequately to describe a balance for each variable.
- Tie the intuitions in the 2-simplex to their graphical representations in the multisplit entropy triangle.
- Provide evidence that the new entropy decomposition and associated representation bring insight into multivariate sources.

We will call this solution the *source multivariate entropy balance equation* and the *source multivariate entropy triangle, (SMET)* and we will refer to the previous solution, described in the sections above, as the *channel bivariate entropy balance equations and triangle*.

We foresee the following problems related to obtaining the balance equation and SMET:

Problem 1. *There is no clear analogue of Mutual Information for multivariate distributions as a candidate to represent information shared by multiple variables to be able to generalize the entropy balance equation and triangle of Sections 2.1 and 2.2.*

This problem is addressed in Section 3.1.

Problem 2. *The entropies of a source and a channel are conceptually different and need different intuitions about the new coordinates.*

This problem is addressed in Section 3.2.

Problem 3. *There is no guarantee that the procedure for splitting the balance equation and triangle in Section 2.4 generalizes to more than two variables.*

This problem is addressed in Section 3.1.

Problem 4. *Related to Problems 2 and 3 above, reading entropies off a source multivariate split entropic triangle does not engender the same intuitions as a channel triangle.*

This problem is addressed in Section 3.3.

Problem 5. *Finding the adequate applications for the new balance equations and triangle representations.*

This problem is addressed in Section 4.

Next, we present our main theoretical results for the generalizations of the balance equation and the entropy triangle for multivariate distributions that we believe solve the aforementioned problems.

3. Theoretical results

In this section we generalize the results presented in Section 2 to the multivariate setting, first on an aggregate basis and then individually for each feature. We will see that these two decompositions provide different insights into the source entropy of a random vector $\overline{X}$.

We start with the balance equation and then we go on to the entropy triangle.

3.1. The source multivariate balance equation

In the context of the random vector $\overline{X} \sim P_{\overline{X}}$, let $\Pi_{\overline{X}} = \prod_{i=1}^{n} P_{X_i}$ be the (jointly) independent distribution with similar marginals to $P_{\overline{X}}$ and $U_{\overline{X}} = \prod_{i=1}^{n} U_{X_i}$ be the uniform distribution with identical support. To highlight the divergence with the uniform multivariate distribution, we introduce:

$$\Delta H_{\Pi_{\overline{X}}} = H_{U_{\overline{X}}} - H_{\Pi_{\overline{X}}} \tag{15}$$

Notice that since uniform distributions are independent, from (15) we may write:

$$\Delta H_{\Pi_{\overline{X}}} = \sum_{i=1}^{n} H_{U_{X_i}} - \sum_{i=1}^{n} H_{P_{X_i}} = \sum_{i=1}^{n} H_{U_{X_i}} - H_{P_{X_i}} = \sum_{i=1}^{n} \Delta H_{P_{X_i}} \tag{16}$$

where $\Delta H_{P_{X_i}} = H_{U_{X_i}} - H_{P_{X_i}}$ defines the divergence of each of the component random variables.

Both the divergence from uniformity and the variation of information have readily available interpretations as areas in the generalization of Figure 3 that is represented in Figure 5, as the yellow and red areas, respectively (Reza, 1961, §3.13). The unaccounted for area is clearly the bound information represented in green in Figure 5.

$$M_{P_{\overline{X}}} = C_{P_{\overline{X}}} + D_{P_{\overline{X}}} = (H_{\Pi_{\overline{X}}} - H_{P\overline{X}}) + (H_{P\overline{X}} - VI_{P_{\overline{X}}}) = H_{\Pi_{\overline{X}}} - VI_{P_{\overline{X}}} \tag{17}$$

Notice that this adds another direction of generalization of the mutual information: in the same way that in the bivariate case we have:

$$H_{P_X P_Y} - VI_{P_{XY}} = M_{P_{XY}} = 2 * MI_{P_{XY}} \tag{18}$$

we now have for the multivariate case a further motivation for the bound information, viz.

$$H_{P_{\overline{X}}} - VI_{P_{\overline{X}}} = M_{P_{\overline{X}}} = C_{P_{\overline{X}}} + D_{P_{\overline{X}}} \tag{19}$$

Through this property, $M_{P_{\overline{X}}}$ may be written in terms of the component entropies:

$$M_{P_{\overline{X}}} = \sum_{i=1}^{n} H_{P_{X_i}} - \sum_{i=1}^{n} H_{P_{X_i|X_i^c}} = \sum_{i=1}^{n} (H_{P_{X_i}} - H_{P_{X_i|X_i^c}}) \tag{20}$$

and let us call $M_{P_{X_i}} = H_{P_{X_i}} - H_{P_{X_i|X_i^c}}$, the *bound information (of X_i)*, the amount of entropy of P_{X_i} that is bound through dependences to the marginal distributions of different orders of $P_{X_i^c}$. Recall that this must be the mutual information between X_i and the rest of the variables in $\overline{X}$, $M_{P_{X_i}} = MI_{X_i X_i^c}$, but this cannot be equal to any of the quantities described in Section 2.5.

Note how all the previously considered quantities are reducible to those about their component variables, a situation that is not too clear in Figure 5. In fact, it will prove very useful later to consider the following conditions for a given variable X_i in the context of $\overline{X}$:

- *Uniformity*, $P_{X_i} = U_{X_i}$, whence $H_{P_{X_i}} = H_{U_{X_i}}$ is maximal with $\Delta H_{P_{X_i}} = 0$. The opposite of this property is *determinacy* whereby $P_{X_i}(x) =$

$\delta_{a_i}(x)$, in which case there is no uncertainty about the outcome of X_i, $H_{P_{X_i}} = 0$, and $\Delta H_{P_{X_i}} = H_{U_{X_i}}$ whence we may conclude:

$$0 = \Delta H_{P_{X_i}|P_{X_i}=U_{X_i}} \leq \Delta H_{P_{X_i}} \leq H_{U_{X_i}} = \Delta H_{P_{X_i}|P_{X_i}=\delta_{a_i}} \tag{21}$$

- *Orthogonality*, $X_i \perp X_i^c$, defined by $P_{\overline{X}} = P_{X_i} P_{X_i^c}$, whence $H_{P_{\overline{X}}} = H_{P_{X_i^c}} + H_{P_{X_i}}$. In such case, since $H_{P_{\overline{X}}} = H_{P_{X_i^c}} + H_{P_{X_i|X_i^c}}$, we conclude that $H_{P_{X_i|X_i^c}} = H_{P_{X_i}}$ and $M_{P_{X_i}} = 0$ by definition.
- *Redundancy*, $X_i \subseteq X_i^c$ if the value of X_i is completely determined by the value of X_i^c . This entails that $H_{P_{X_i|X_i^c}} = 0$.

As a result, we see that there are bounded continua for the values of $H_{P_{X_i|X_i^c}}$ and $M_{P_{X_i}}$

$$H_{P_{X_i|X_i^c}|X_i \subseteq X_i^c} \equiv 0 \leq H_{P_{X_i|X_i^c}} \leq H_{P_{X_i}} \equiv H_{P_{X_i|X_i^c}|X_i \perp X_i^c} \tag{22}$$

$$M_{P_{X_i}|X_i \perp X_i^c} \equiv 0 \leq M_{P_{X_i}} \leq H_{P_{X_i}} \equiv M_{P_{X_i}|X_i \subseteq X_i^c} \tag{23}$$

For the reasons above, in order to solve Problem 1 we propose to use the bound information of $\overline{X}$, as the third coordinate in a balance equation, since in the new variables it is easy to write a new balance equation.

Theorem 1 (Aggregate source multivariate balance equation)**.** *Let $P_{\overline{X}}$ be an arbitrary discrete distribution over the set of random variables $\overline{X}$. Then, with the definitions above, the following* balance equation holds

$$\begin{aligned} H_{U_{\overline{X}}} &= \Delta H_{\Pi_{\overline{X}}} + M_{P_{\overline{X}}} + VI_{P_{\overline{X}}} \\ 0 &\leq \Delta H_{\Pi_{\overline{X}}}, M_{P_{\overline{X}}}, VI_{P_{\overline{X}}} \leq H_{U_{\overline{X}}} \end{aligned} \tag{24}$$

Proof. To prove the balance equation add together (15) and (17) and reorganize.

Regarding the bounds, for those of the divergence from uniformity consider (16) and the inequalities (21). Since the individual divergences are all non-negative we may add to obtain

$$0 \leq \Delta H_{\Pi_{\overline{X}}} = \sum_{i=1}^{n} \Delta H_{P_{X_i}} \leq \sum_{i=1}^{n} H_{U_{X_i}} = H_{U_{\overline{X}}} \tag{25}$$

By (7) and (22) we see that the variation of information $VI_{P_{\overline{X}}}$ is a sum of nonnegative quantities. If each of the component distributions is uniform, $\forall i,\ P_{X_i} = U_{X_i}$, we have:

$$0 = \sum_{i=1}^{n} H_{P_{X_i|X_i^c}} \underset{|X_i \subseteq X_i^c}{\leq} VI_{P_{\overline{X}}} = \sum_{i=1}^{n} H_{P_{X_i|X_i^c}} \leq H_{U_{\overline{X}}} = \sum_{i=1}^{n} H_{P_{X_i}|P_{X_i}=U_{X_i}} \tag{26}$$

Note that $\sum_{i=1}^{n} H_{P_{X_i}|P_{X_i}=U_{X_i}} = \sum_{i=1}^{n} H_{P_{X_i|X_i^c}|X_i \perp X_i^c \& P_{X_i}=U_{X_i}}$. Similarly, by (20) and (23), for the bound information when each of the components is uniform,

$$0 = \sum_{i=1}^{n} M_{P_{X_i}} \underset{|X_i \perp X_i^c}{\leq} M_{P_{\overline{X}}} = \sum_{i=1}^{n} M_{P_{X_i}} \leq H_{U_{\overline{X}}} = \sum_{i=1}^{n} M_{P_{X_i}|P_{X_i}=U_{X_i}} \tag{27}$$

And $\sum_{i=1}^{n} M_{P_{X_i}|P_{X_i}=U_{X_i}} = \sum_{i=1}^{n} M_{P_{X_i}|X_i \subseteq X_i^c \& P_{X_i}=U_{X_i}}$. □

We believe that in the proof of Theorem 1 is the solution to Problem 1, but in fact, we have also proven a more restrictive theorem that solves Problem 3.

Theorem 2 (Multi-split source multivariate balance equation)**.** *Let $P_{\overline{X}}$ be an arbitrary discrete distribution over the set of random variables $\overline{X} = \{X_i\}_{i=1}^{n}$. Then, with the definitions above, the* balance equation holds *for each variable individually:*

$$\begin{aligned} H_{U_{X_i}} &= \Delta H_{P_{X_i}} + M_{P_{X_i}} + H_{P_{X_i|X_i^c}}, \quad 1 \leq i \leq n \\ 0 &\leq \Delta H_{P_{X_i}}, M_{P_{X_i}}, H_{P_{X_i|X_i^c}} \leq H_{U_i}, \quad 1 \leq i \leq n \end{aligned} \tag{28}$$

Proof. We notice that:

$$H_{U_{X_i}} = (H_{U_{X_i}} - H_{P_{X_i}}) + (H_{P_{X_i}} - H_{P_{X_i|X_i^c}}) + H_{P_{X_i|X_i^c}}$$

and then we identify the terms with the names defined above. The inequalities were proven in (21) and, *a fortiori*, by (26) and (27) in the proof of Theorem 1. □

Notice that under this new, more detailed point of view, the aggregate balance equation (24) is just the addition of the individual (28) over all random variables. In order to distinguish between them, we call (1) and all results issuing from it *aggregate*, while we will call (28) and all results issuing from it *multisplit*.

3.2. The aggregate source multivariate entropy triangle

Similarly to the procedure (2), we may normalize the aggregate source multivariate balance equation by $H_{U_{\overline{X}}}$ to obtain

$$\begin{aligned} 1 &= \Delta H'_{\Pi_{\overline{X}}} + M'_{P_{\overline{X}}} + VI'_{P_{\overline{X}}} \\ 0 &\leq \Delta H'_{\Pi_{\overline{X}}}, M'_{P_{\overline{X}}}, VI'_{P_{\overline{X}}} \leq 1 \end{aligned} \tag{29}$$

The composition $F(P_{\overline{X}}) = [\Delta H'_{P_{\overline{X}}}, M'_{P_{\overline{X}}}, VI'_{P_{\overline{X}}}]$ suggests a representation in terms of a ternary diagram on the aggregate entropy with similar meaning as before:

- If $P_{\overline{X}} = \Pi_{\overline{X}} = \Pi_{i=1}^{n} P_{X_i}$ then $F(P_{\overline{X}}) = [\cdot, 0, \cdot]$, is the geometric locus of distributions with independent marginals and has a high residual entropy.

- If $P_{X_i} = U_{X_i}, 1 \leq i \leq n$ then $F(P_{\overline{X}}) = [0, \cdot, \cdot]$ is the geometric locus of distributions with uniform marginals.

- If $P_{X_i} = P_{X_j}, i \neq j$ then $F(P_{\overline{X}}) = [\cdot, \cdot, 0]$ is the locus of distributions with identical marginals and in general high bound information.

However, the interpretations of the variables of the SMET are very different to what they were in the channel triangle:

1. For instance, the multivariate residual entropy $VI_{P_{\overline{X}}}$ is actually the sum of amounts of information singularly captured by each variable. Nowhere else can it be found and any later processing that ignores this quantity will incur in the deletion of that information, e.g. for transmission purposes.

2. Likewise, the total bound information is highly redundant in that every portion of it resides in (at least two) different variables. Once the entropy of one feature has been processed, the part of the bound information that lies in it is redundant for further processing.

3. Somewhat similar to the original interpretation, the divergence from uniformity is not available for processing. It is a potentiality—maximal randomness—of the source of information that has not been realized and therefore is *not available for later processing*, unlike the other entropies are.

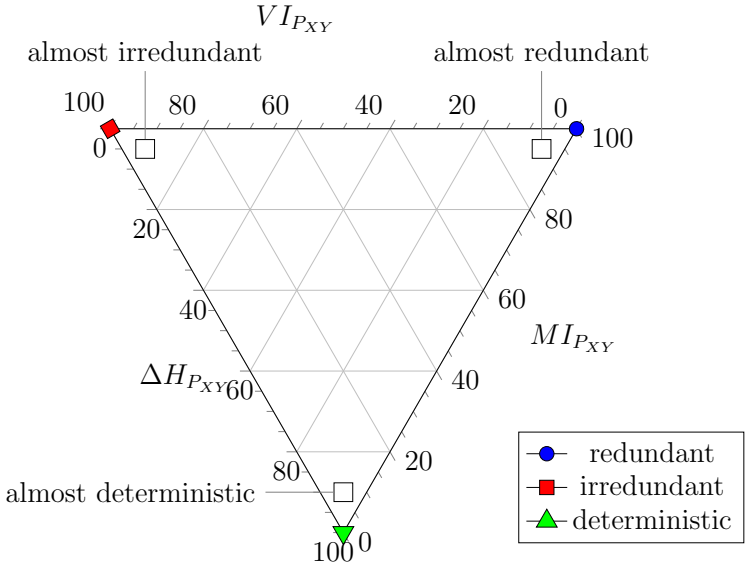


Figure 6: Conceptually annotated Source Multivariate Entropy Triangle

Since this latter quantity is deleterious to information transmission, a different representation to that of the original triangle suggests itself: *the simplex should be rotated so that the divergence from uniformity is represented as a down-growing quantity*. The rationale for this is that the lower a distributtion is plotted, the less information it has at its disposal to be transmitted. This representation is what we call the *aggregate Source Multivariate Entropy Triangle (SMET)*.

Figure 6 shows a conceptual version of the SMET annotated with these intuitions. Although this might not be the whole meaning of the three coordinates, we believe that this solves Problem 2.

3.3. The multisplit source multivariate entropy triangle

The aggregate visualization and information structure of a joint distributions is an average of the component decompositions for the different variables, as implied by (24)-(28) . A finer, disaggregate visualization and analysis tool is to be introduced next.

It is clear that the normalization carried out for (2) and (20) can also be carried for the split balance equations (28), but in this case we use each of

the individual $H_{U_{X_i}}$ to obtain:

$$\begin{aligned} 1 &= \Delta H'_{P_{X_i}} + M'_{P_{X_i}} + H'_{P_{X_i|X_i^c}}, \quad 1 \leq i \leq n \\ 0 &\leq \Delta H'_{P_{X_i}}, M'_{P_{X_i}}, H'_{P_{X_i|X_i^c}} \leq 1 \end{aligned} \tag{30}$$

Then for each multivariate $\overline{X} = \{X_i\}_{i=1}^n$ we may write for each marginal P_{X_i} the coordinates in a de Finetti diagram as $F(P_{X_i}) = [\Delta H'_{P_{X_i}}, M'_{P_{X_i}}, H'_{P_{X_i|X_i^c}}]$, *with similar interpretation as in Section 3.2 but regarding the content of a single variable.*

Notice that despite the fact that these coordinates all refer to potentially different entropy levels—since the normalizig $H_{U_{X_i}}$ may vary greatly for each X_i—in the normalized form they can all be represented in the same entropy triangle. We refer to this common representation as the *multisplit Source Multivariate Entropy triangle (multisplit SMET).*

With this new arrangement in place, the upper right-hand angle of the inverted triangle represents the locus of *highly redundant variables*, whereas the left-hand angle represents that of *highly irredundant variables* with an extensive amount of information that only pertains to them. Finally, the lower angle in the triangle represents almost deterministic variables, conveying very little information in general.

As an example, Figure 7 shows a multisplit SMET annotated with distributions obtained by different means:

- Several *irredundant* distributions obtained as binomial distributions of m instances with parameter $p = 0.5$. We expect these distributions to be pairwise independent $H'_{P_{X_i|X_i^C}} \approx 1$, that is, to lay close to the upper left-hand angle.

- Several almost *deterministic* distributions obtained as binomial distributions of m instances where $p = 0.99$. We expect these distributions to have $\Delta H'_{X_i} \approx 1$ that is, to lay close to the bottom angle, where there is no irredundant nor any bound information

- Several distributions which are a binomial distribution with parameter $p = 0.5$ with m realizations of a small noise added. We can see that these contain a lot of bound information, that is, are highly *redundant* as $M'_{P_{X_i}} \approx 1$, and lie close to the upper right-hand angle.

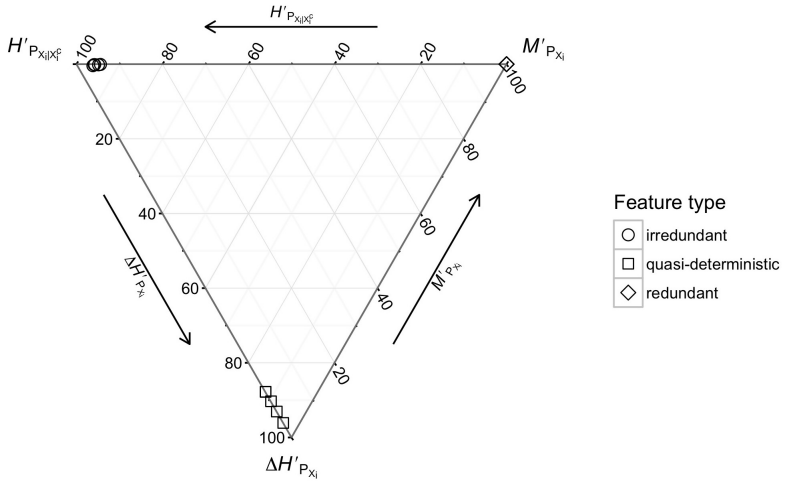


Figure 7: **Multisplit Source Multivariate Entropy Triangle (SMET) of three contrived, extreme datasets.** On the left upper corner, a dataset with four irredundant variables; on the right upper corner, one with a binomial distribution to which four different noise realizations were added; on the bottom, one with four quasi-deterministic features.

Finally, notice that in the light of the comment at the end of the previous Section, we can also represent the quantity developed in 3.2—the aggregate of all the effects of the individual marginals—in this multisplit triangle. Triangles of this sort will be shown in Section 4.

Under these interpretations, we believe that Problem 4 is solved. Example applications of these theoretical results will be explained next.

4. Example Applications

The SMET is intended as an exploratory analysis tool (Tukey, 1977), hence the envisaged applications will be posed as *analysis questions*, specifically on the information content of datasets, and in doing so will provide a solution to Problem 5 in the belief that many more applications will follow. For instance, a more informed choice of clustering algorithms for each dataset could be made after this analysis stage. Other possibilities are explored in Section 5.

4.1. Characterizing the information content of machine learning tasks

We first tackle the problem of characterizing the information content of machine learning task datasets. The question to be answered here is *what is*

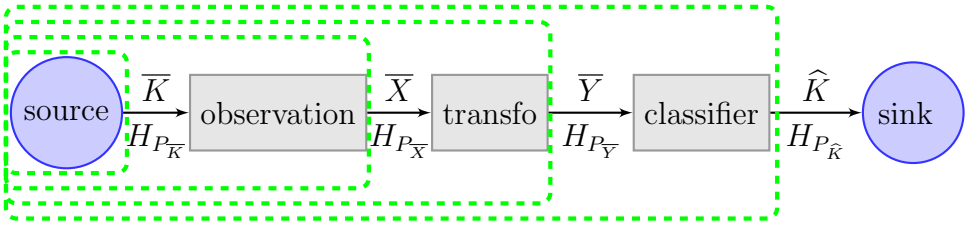


Figure 8: **Schematic representation of a multi-class/multi-label classification task with points where the SMET is applicable**. The dotted boxes represent measuring points where the abstraction of a multivariate source could be of interest: for investigating the classification labels themselves $\overline{K}$, the features as issued from a process of observation $\overline{X}$, as observed after some feature transformation $\overline{Y}$, or even the result of classification $\widehat{K}$.

the information composition of the data in this task?

Figure 8 gives a schematic of a machine learning task for the case of *supervised* multi-class/multi-label classification and annotated with the entropies of all variables involved, unlike Figure 1.(c). But if we consider that $\widehat{K}$ are not available or just ignore them, the diagram in Figure 8 covers also unsupervised tasks.

We can use the framework developed in this paper to analyze the information content of any of the virtual data sources represented by the dashed rectangles, as instances of those systems initially introduced in Figure 2. Specifically, we may study statistical multivariate sources $\overline{K}$, the result of multivariate observation processes on those sources $\overline{X}$, or the result of multivariate transformation on the observations $\overline{Y}$. In this paper, purposely, we do not study the transformation of $\overline{X}$ into $\overline{Y}$ or its results.

For the purpose at hand, we have used the R environment (R Core Team, 2015) and databases from the `vcd` (Meyer et al., 2015) and `mlbench` (Leisch & Dimitriadou, 2010) packages, some of which belong to the UCI repository (Lichman, 2013). The visualizations are done using our own R package[3], but there are other resources for compositional data available (van den Boogaart & Tolosana-Delgado, 2013; Hamilton, 2015).

To use the same data other applications below, instead of selecting unsupervised (clustering) data, we selected supervised multi-class classification data for the `Arthritis`, `iris`, `Ionosphere`, `Glass` and `BreastCancer`

[3] `https://github.com/FJValverde/entropies.git`

databases since they are widespread and practitioners already have their own intuition as to whether our inferences are likely or not. In the diagrams of Figure 2 this corresponds to considering a model of process observation like that of Figure 2.(c), rather than the unsupervised model of Figure 2.(b). Table 1 lists their more evident characteristics

	Dataset Name	$\lvert\text{support}(K)\rvert$	$\lvert\overline{X}\rvert$	instances
1	Ionosphere	2	34	351
2	iris	3	4	150
3	Glass	7	9	214
4	Arthritis	3	3	84
5	BreastCancer	2	9	699
6	Sonar	2	60	208
7	Wine	3	13	178

Table 1: Some datasets considered in this study

Figure 9 shows the aggregate entropy characterization of the datasets above. We can clearly see that `Arthritis` is mostly composed of irredundant features, while `Ionosphere`, `Glass`, `Sonar` and `Wine` are almost entirely composed of a redundant set of features. The features of `Iris` and `BreastCancer` are in an intermediate level of redundancy. Since all of the datasets but `BreastCancer` are almost perfectly balanced, the redundancy is caused by the information of each featured being bound to a combination of others. On the other hand, the difference in balance between `Iris` (perfect) and `BreastCancer` is quite marked. These differences are explored further below.

4.2. How do features contribute to the aggregate information?

A natural question to raise next would be to see how the different features are represented in each of these datasets depicted in Figure 9. The question being posed here is *what are the compositions or balances of individual features in the dataset?*

This can be solved with the entropy decomposition of (28) and the multisplit SMET of Section 3.3. For instance, Figure 10 shows this behavior for two different tasks:

- For `iris` there is a variety of behaviors with features ranging from sligthly irredundant to rather redundant. The aggregate, consequently, tends to the latter.

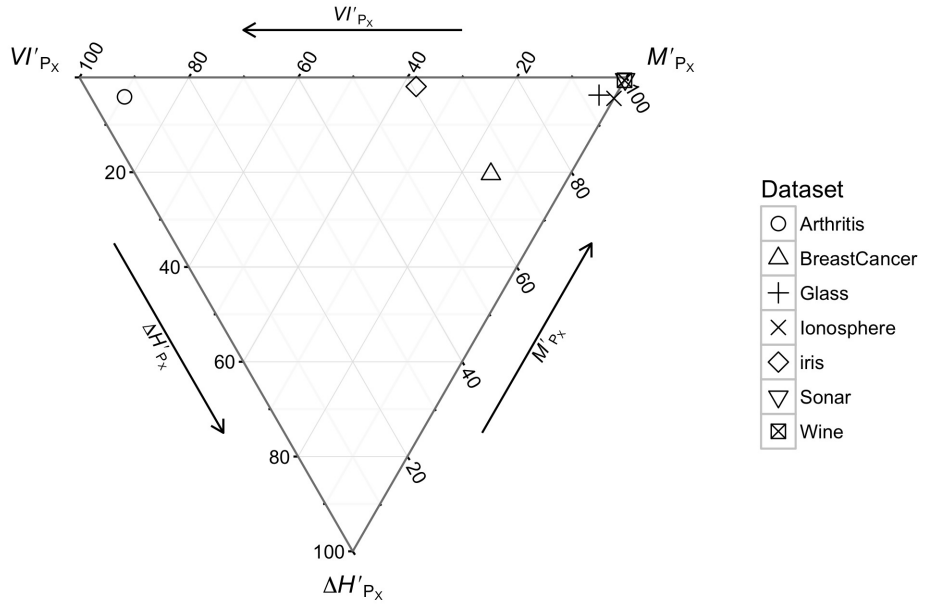


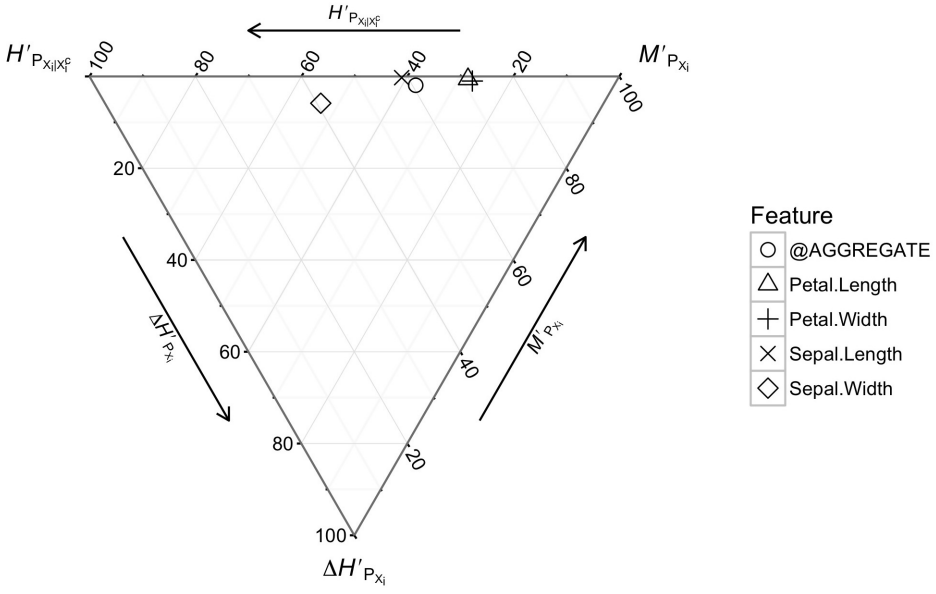
Figure 9: **Aggregated Multivariate Source Entropy Triangle for some datasets.** Clearly `Arthritis`, on one hand, `Iris` and `BreastCancer`, on another, and the rest of the datasets belong to three different types of datasets.

- For `Glass` most of the features are very redundant, all of their information bound with some other. Furthermore, two features, `Fe` and `Ba` are quite unbalanced, providing less bound information than the rest. Overall the aggregate is balanced and redundant.

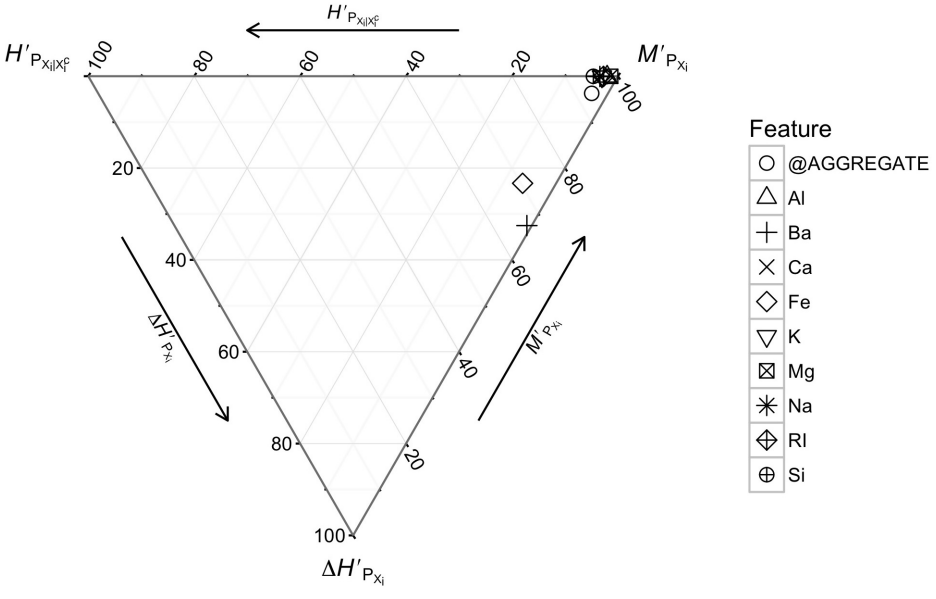
Although this is all of the information that may be explored in the unsupervised case, we can explore further in the supervised case, as in the section below.

4.3. How well do features encode the information in the class?

When considering supervised data, an issue often ignored is whether the data, including the class variable, are adequate to undergo machine learning procedures. The kind of questions we will try to answer are of the type a*re the data in the task good enough to solve it?*



(a) Dataset: `iris`



(b) Dataset: `Glass`

Figure 10: **Multisplit Source Multivariate Entropy Triangles for the `iris` and `Glass` data.** The aggregate entropies as in the previous Figure are under label `@AGGREGATE`. Their most notorious difference lies in their balance and redundancy.

We can try to answer this question by studying the balance equation of the class variable in the context of the dataset. Consider the class variable is, without loss of generality, the first feature of the dataset $K = X_o$. Then by instantiating (28) and (30) we have

$$\begin{aligned} H_{U_K} &= \Delta H_{P_K} + M_{P_K} + H_{P_{K|K^c}} \\ 1 &= \Delta H'_{P_K} + M'_{P_K} + H'_{P_{K|K^c}} \end{aligned}$$

where K^c represents the features of the dataset. We interpret these as:

- ΔH_{P_K} quantifies how unbalanced the class variable is. From (Valverde-Albacete & Peláez-Moreno, 2010) we know that the greater this quantity, the easier determining it with any possible encoding. In the triangle, the further down the $\Delta H'_{P_K}$ quantity the easier K is to determine from any feature.

- M_{P_K} quantifies the information of K provided by other variables K^c. The higher M'_{P_K} (the more to the right) the more quantity of information in K is captured by the features and the easier the classification task is.

- $H_{P_{K|K^c}}$ is therefore the remanent entropy in the class variable not captured by the features. Consequently, the higher $H'_{P_{K|K^c}}$ in the ET (the more to the left) the more difficult the classification task is.

Thus the position in the triangle for the normalized, multisplit balance equation provides information as to the difficulty of the classification task, prior to any inference of a classifier.

As an example of this, Figure 11 shows the plot of the split entropies for the class labels in the datasets used in the previous sections. We can see three kind of datasets in it:

- balanced datasets—`iris`, `Sonar` and `Wine`—that have no irredundant information in their class variable. These can be solved perfectly with 100% (Entropy-Modified) Accuracy (Valverde-Albacete & Peláez-Moreno, 2014).

- *almost* balanced datasets with no irredundant information—`Breastcancer`, `Glass` and `Ionosphere`. These can be solved, but not perfectly: a 100% accuracy would mean that some information that is not in the features has "magically" been supplied by the classifier.

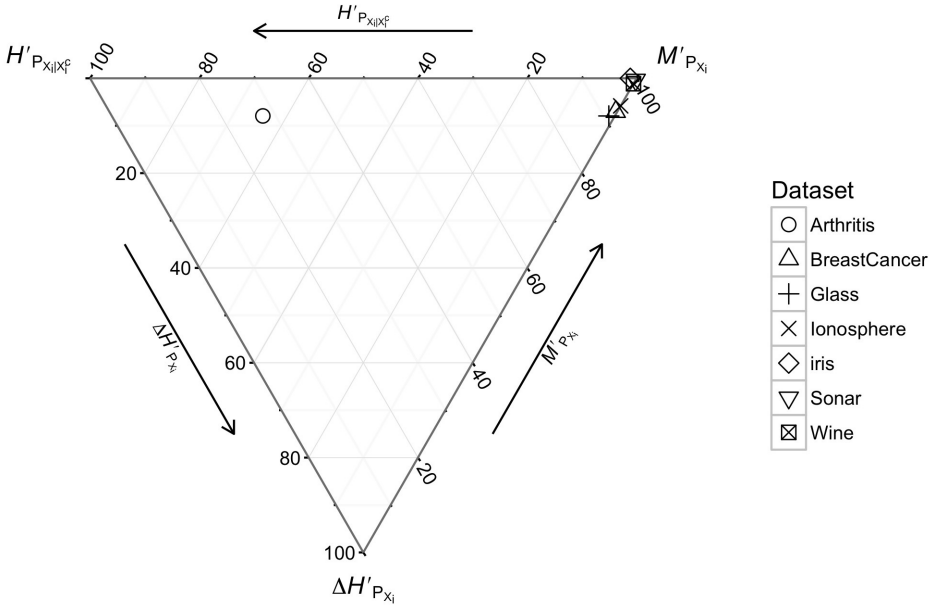


Figure 11: **Class label composition shift for the datasets considered.** The balance of the class is only indirectly related to the balancedness of the dataset. The irredundant information of the labels quantifies in a straightforward way how hard the classification task is.

- one almost balanced dataset with a lot of irredundant information—`Arthritis` [4]—not captured by the features and therefore this new task is more difficult than those above: no effectiveness can be attained in this task above a limit, as yet unexplored.

The next section introduces a discussion on alternate representations for entropy balances and other possible applications of the SMET.

5. Discussion

5.1. A Stacked Chart for Source Entropies

Plotting the entropic decomposition of a source in the entropy triangle entails and implicit normalization that is not satisfactory at times: e.g. when

[4]This database was originally conceived to establish whether a given medical treatment was significatively more effective than a placebo to attenuate the symptoms of arthritis patients. It is worth noting that we are using it in a classification setting and therefore the task is now to obtain the degree of improvement (the class variable) from the features.

measuring the absolute levels of entropy related to a particular dataset and its variables. Note that difficulty pertains to all compositional data and not just the decomposition being made evident in the SMET. In this section we want to discuss whether there is an alternative to the Entropy Triangle for entropic compositional data.

For instance, we could use a stacked bar graph with entropy bars for each variable (and possible the whole source itself.) Figure 12 presents the decomposition of the source multivariate entropy of two differents tasks by their absolute values. Since the interesting quantities to consider are the remaining entropies and the dual total correlation, it is often cognitively useful to print the variables in terms of descending total $M_{P_{X_i}} + H_{P_{X_i|X_i^C}}$.

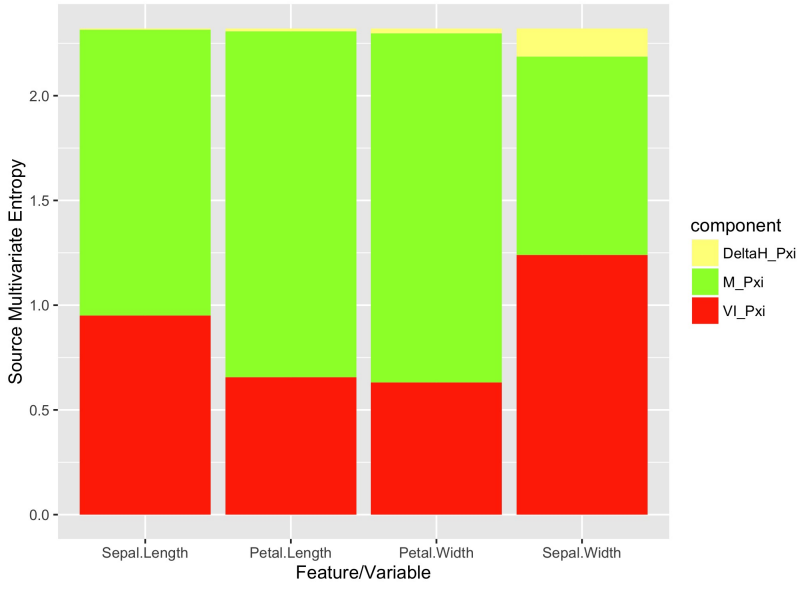
Alternate ways to represent the information contained in stacked bar graphs (Tufte, 1992) like those of Figure 12 are proportional *stacked bar graphs* and *pie charts*.

Regarding the first, stacking bars proportionally essentially does away with the advantage of stacked bars over the entropy triangle, namely, providing the absolute numbers in the decomposition. We believe that it pays the cognitive effort of learning to read an entropy triangle since the information being represented can be much more comprehensive. While the proportional stacked bar graph represents the proportions of the composition of the total entropy, the SMET can use glyphs—as in Figure 9—or any other device to convey more information in the graph.

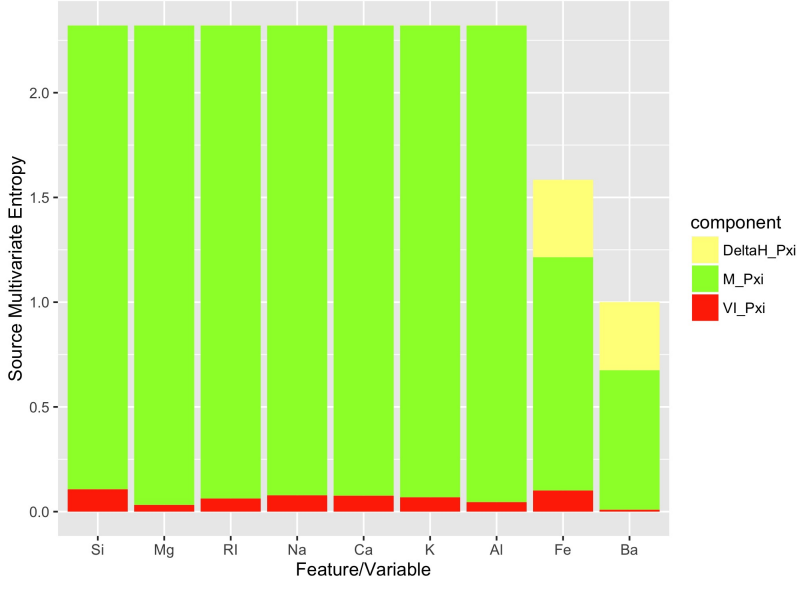
For instance, consider the graphs in Figure 13. Figure 13.(a) is a modified SMET that represents the absolute value of each variable's total entropy $H_{U_{X_i}}$ using a sidebar. Compare it to Figure 12.(b): perhaps the direct comparison of total entropies is difficult to glean from the former, but the intuitions stemming from the coordinates in the SMET are completely missing from the latter. We foresee that this type of diagrams will become more interesting to the user as she gets used to reading the information from the SMET axes.

While Figure 13.(b) encodes the same information as Figure 13.(a), humans are notorious at misreading information encoded in areas, as in a pie chart. Indeed, comparing it to Figure 12.(b) we can see how the remaining entropy $H_{P_{X_i}}$ is mis-represented as a small angular sector for a number of variables in the pie charts.

We conclude that the entropy triangle has advantages over both stacked bar and pie charts, at the expense of learning a new type of chart, if the user is previously unacquainted with ternary diagrams. Otherwise, we believe the

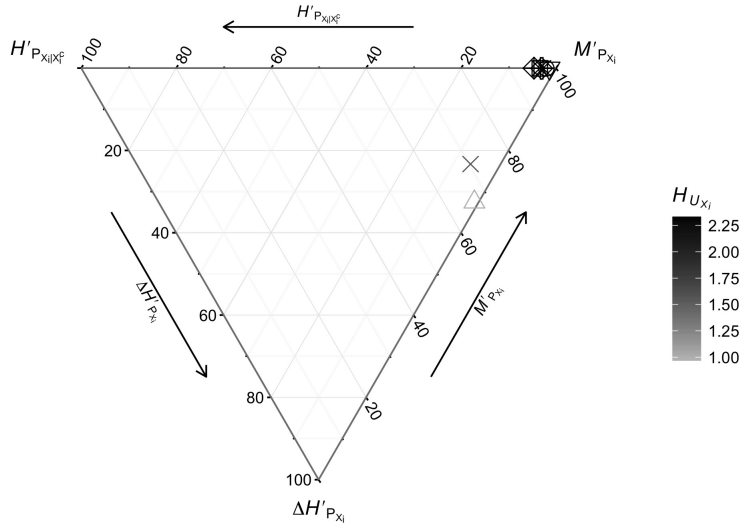


(a) Dataset: `iris`

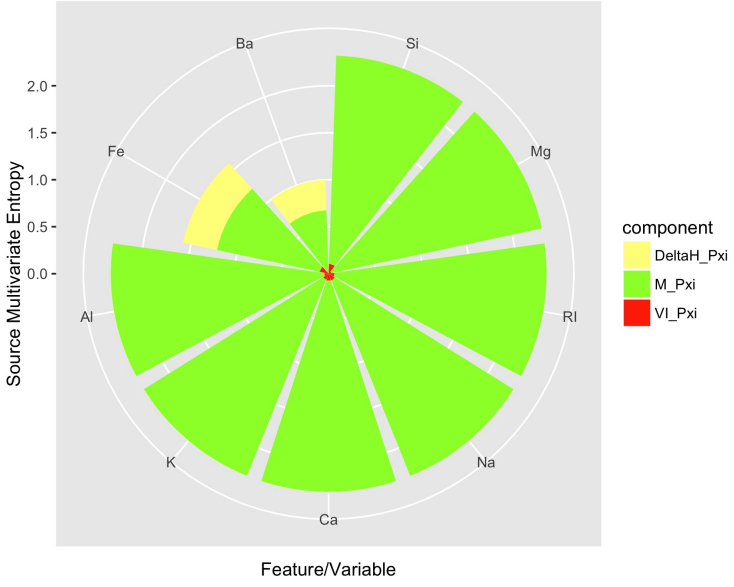


(b) Dataset: `Glass`

Figure 12: **(Color online) Entropy decomposition of *iris* (above) and *Glass* (below) into its components for each variable.** These values correspond to those with the glyphs in Fig. 10, and the color of the bars correspond to those of Fig. 5.



(a) Entropy triangle with colorbar



(b) Stacked pie chart

Figure 13: **Color online) Entropy decomposition of *Glass* into its components for each variable using the SMET (above) and a pie chart (below).** The glyphs are the same as in Figure 10.(b).

advantages of using the SMET for entropy description greatly surpass the learning period required to adapt to them.

5.2. Other possible applications

Care has to be taken not to misconstrue the application in Section 4.3 for a feature selection process. Indeed, the existence of a right shift in the entropy structure caused by the class variable can be taken to indicate a strong informational bound to all shifted features. But the accumulation of features per se does not guarantee that much information *about the class* is available. For that purpose the information conditioned on the class variable should be considered instead (Brown et al., 2012).

In practice, what this amount to saying is that feature selection needs a *channel* multivariate balance equation and entropy triangle: much like end-to-end evaluation needs the bivariate versions appropriate for Figures 1.(b) and 2.(a), feature selection needs the channel versions of the SMET and balance equations. This is left for future work.

A very suggestive use of these multivariate tools would be to explore how the performance of different classifier induction schemes—e.g. NN, tree classifiers, logistic regressors, etc.—varies with datasets which have quantitatively distinct positions in the SMET. This would call for extensive experimental work that would surely exceed the length of a paper where the techniques are first introduced, like the present one.

6. Conclusions

We have found an informative and promising description and representation for the entropic content of multivariate distributions, casting it as an instance of compositional data, that deserves better exploration in further work.

We have also demonstrated its use for the visualization of the total and split entropy in datasets on a per-feature basis. A brief analysis has shown that it is possible to detect which features contribute the most to the total bound information in a classification dataset, which is one of the "substances" whose transmission is to be "maximized" in classification tasks.

We believe that the combination of this representation with that already introduced in previous papers will allow us to explore how information is transmitted from the data in the datasets to the intelligence represented by

classifiers in machine learning tasks, including multi-class and multi-label classification. This is left for future work.

Finally, some conclusions will be outlined in the next section.

7. Acknowledgments

CPM & FVA have been partially supported by the Spanish Government-MinECo projects TEC2014-53390-P and TEC2014-61729-EXP in this research.

Abdallah, S. A. & Plumbley, M. D. (2010). *Predictive Information, Multi-information and Binding Information*. Technical Report C4DM-TR10-10, Queen Mary, University of London.

Abdallah, S. A. & Plumbley, M. D. (2012). A measure of statistical complexity based on predictive information with application to finite spin systems. *Physics Letters A*, 376(4), 275–281.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44(2), 139–177.

Bell, A. (2003). The co-information lattice. In N. Murata, S.-i. Amari, A. Cichocki, & S. Makino (Eds.), *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation*.

Brillouin, L. (1962). *Science and information theory*. Second edition. Academic Press, Inc., Publishers, New York.

Brown, G., Pocock, A., Zhao, M.-J., & Luján, M. (2012). Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *Journal of Machine Learning Research*, (pp. 27–66).

Chen, T., Jin, Y., Qiu, X., & Chen, X. (2014). A hybrid fuzzy evaluation method for safety assessment of food-waste feed based on entropy and the analytic hierarchy process methods. *Expert Systems with Applications*, 41(16), 7328 – 7337.

Gibaja, E. & Ventura, S. (2015). A Tutorial on Multilabel Learning. *ACM Computing Surveys (CSUR)*, 47(3), 52–38.

Hamilton, N. (2015). *ggtern: An Extension to ggplot2, for the Creation of Ternary Diagrams*. R package version 1.0.6.1.

Han, T. S. (1978). Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36(2), 133–156.

Hempelmann, C. F., Sakoglu, U., Gurupur, V. P., & Jampana, S. (2016). An entropy-based evaluation method for knowledge bases of medical information systems. *Expert Systems with Applications*, 46, 262 – 273.

James, R. G., Ellison, C. J., & Crutchfield, J. P. (2011). Anatomy of a bit: Information in a time series observation. *Chaos*, 21(3), 037109–037109.

Jaynes, E. T. (1996). *Probability theory: The logic of science*. Cambridge University Press.

Leisch, F. & Dimitriadou, E. (2010). *mlbench: Machine Learning Benchmark Problems*. R package version 2.1-1.

Lichman, M. (2013). UCI machine learning repository.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, 19(2), 97–116.

Meila, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 28, 875–893.

Mejía-Navarrete, D., Gallardo-Antolín, A., Peláez-Moreno, C., & Valverde-Albacete, F. J. (2011). Feature extraction assessment for an acoustic-event classification task using the entropy triangle. In *Interspeech 2010: 12th Annual Conference of the International Speech Communication Association*.

Meyer, D., Zeileis, A., , & Hornik, K. (2015). *vcd: Visualizing Categorical Data*. R package version 1.4-1.

Murphy, K. P. (2012). *Machine Learning*. A Probabilistic Perspective. MIT Press.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Statistics in practice. Chichester, UK: John Wiley & Sons.

Principe, J. C. (2010). *Information Theoretic Learning*. Information Science and Statistics. New York, NY: Springer New York.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reza, F. M. (1961). *An introduction to information theory*. McGraw-Hill Electrical and Electronic Engineering Series. McGraw-Hill Book Co., Inc., New York-Toronto-London.

Rödder, W., Brenner, D., & Kulmann, F. (2014). Entropy based evaluation of net structures - deployed in social network analysis. *Expert Systems with Applications*, 41(17), 7968 – 7979.

Shannon, C. E. (1948a). A mathematical theory of Communication. *The Bell System Technical Journal*, XXVII(3), 379–423.

Shannon, C. E. (1948b). A mathematical theory of communication. *The Bell System Technical Journal*, XXVII(3), 623–656.

Studený, M. & Vejnarová, J. (1998). The Multiinformation Function as a Tool for Measuring Stochastic Dependence. In *Learning in Graphical Models* (pp. 261–297). Dordrecht: Springer Netherlands.

Sun Han, T. (1980). Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46(1), 26–45.

Theodoridis, S. & Koutroumbas, K. (2006). *Pattern Recognition*. Academic Press, third edition.

Tononi, G. (1998). Complexity and coherency: integrating information in the brain. *Trends in Cognitive Sciences*, 2(12), 474–484.

Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences of the United States of America*, 91(11), 5033–5037.

Tufte, E. R. (1992). *The visual display of quantitative information.* Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesley.

Valverde-Albacete, F. J. (2016). **entropies** – An R package to work with entropic coordinates, the entropy triangle, NIT and EMA.

Valverde-Albacete, F. J., de Albornoz, J. C., & Peláez-Moreno, C. (2013). A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks. In P. Forner, henning Müller, R. Paredes, P. Rosso, & B. Stein (Eds.), *Information Access Evaluation. Multilinguality, Multimodality and Visualization. Proceedings of CLEF 2013*, volume 8138 of *LNCS* (pp. 41–52).: Springer.

Valverde-Albacete, F. J. & Peláez-Moreno, C. (2010). Two information-theoretic tools to assess the performance of multi-class classifiers. *Pattern Recognition Letters*, 31(12), 1665–1671.

Valverde-Albacete, F. J. & Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. *PLOS ONE*.

Valverde Albacete, F. J. & Peláez-Moreno, C. (2016). The multivariate entropy triangle and applications. In *Hybrid Artificial Intelligence Systems. HAIS 2016, Sevilla, Spain, April, 2016. Proceedings* (pp. 647–658).: Springer.

van den Boogaart, K. G. & Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R.* Berlin, Heidelberg: Springer Science & Business Media.

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *International Business Machines Corporation. Journal of Research and Development*, 4(1), 66–82.

Zhou, M., Tian, Z., Xu, K., Yu, X., & Wu, H. (2013). Theoretical entropy assessment of fingerprint-based wi-fi localization accuracy. *Expert Systems with Applications*, 40(15), 6136 – 6149.